Как мы DNS в CDN делали

Константин Новаковский





По-настоящему глобальная CDN

Сверхбыстрая доставка любого тяжёлого контента в любую точку мира

Всемирный хостинг

Надёжные виртуальные и выделенные серверы по суперценам. хранилище

Стриминговая платформа

Поддерживаем все этапы трансляции: от создания и захвата видео до воспроизведения

Глобальная защита от DDoS-атак

Защита от сложных DDoS-атак для серверов и веб-приложений

Объектное хранилище

Облачное хранение контента рядом с клиентами

Производительное облако

Создание и масштабирование виртуальной инфраструктуры в несколько кликов

Управление IT-инфраструктурой

Администрирование веб-сервисов без лишних вложений

Разработка софта

Удалённая разработка полного цикла — от проектирования до внедрения

Тестирование игр и веб-сервисов

Улучшение качества ваших программных продуктов

Быстрый DNS-хостинг

Ускорение работы ваших веб-ресурсов





CDN в цифрах



80+
точек присутствия на 5 континентах



800+
кеш-серверов



5 000+
партнёров по пирингу



50+ Тбит/с
пропускная способность сети



<30 мс
среднее время отклика по миру



3 500+
довольных клиентов



США

Ашберн
Атланта
Чикаго
Даллас
Денвер
Лос-Анджелес
Манассас
Майами
Нью-Йорк
Сан-Хосе
Сиэтл
Саннивейл

Латинская Америка

Богота
Буэнос-Айрес
Сан-Паулу
Сантьяго
Лима
Мехико
Рио де Жанейро

Канада

Торонто

Австралия

Мельбурн
Сидней

Европа

Амстердам
Франкфурт
Лондон
Люксембург
Мадрид
Милан
Париж
Прага
Стокгольм
Варшава
Будапешт
София

Азия

Бишкек
Гонконг
Мумбаи
Сеул
Сингапур
Токио
Бангкок
Тайбэй

MENA

Дубай
Стамбул
Тель-Авив

Россия и СНГ

Аксай
Алматы
Ангарск
Барнаул
Владивосток
Воронеж
Екатеринбург
Казань
Киев
Красноярск
Минск
Москва
Нижний Новгород
Новосибирск
Нур-Султан
Орёл
Павлодар
Пермь
Петрозаводск
Псков
Ростов-на-Дону
Санкт-Петербург
Самара
Ташкент
Уфа
Хабаровск
Челябинск

скоро





Наши клиенты

Среди наших клиентов телеканалы, радио, банки, мобильные операторы, рекламные платформы, ретейлеры, интернет-магазины, беттинг, разработчики и издатели игр.









































































HL
HighLoad++

Как доставляется контент?

Клиентское приложение (браузер) делает 2 вещи:

- Получает IP-адрес сервера через DNS
- Выполняет запрос используя IP-адрес



За что боремся?

Клиентское приложение (браузер) делает 2 вещи:

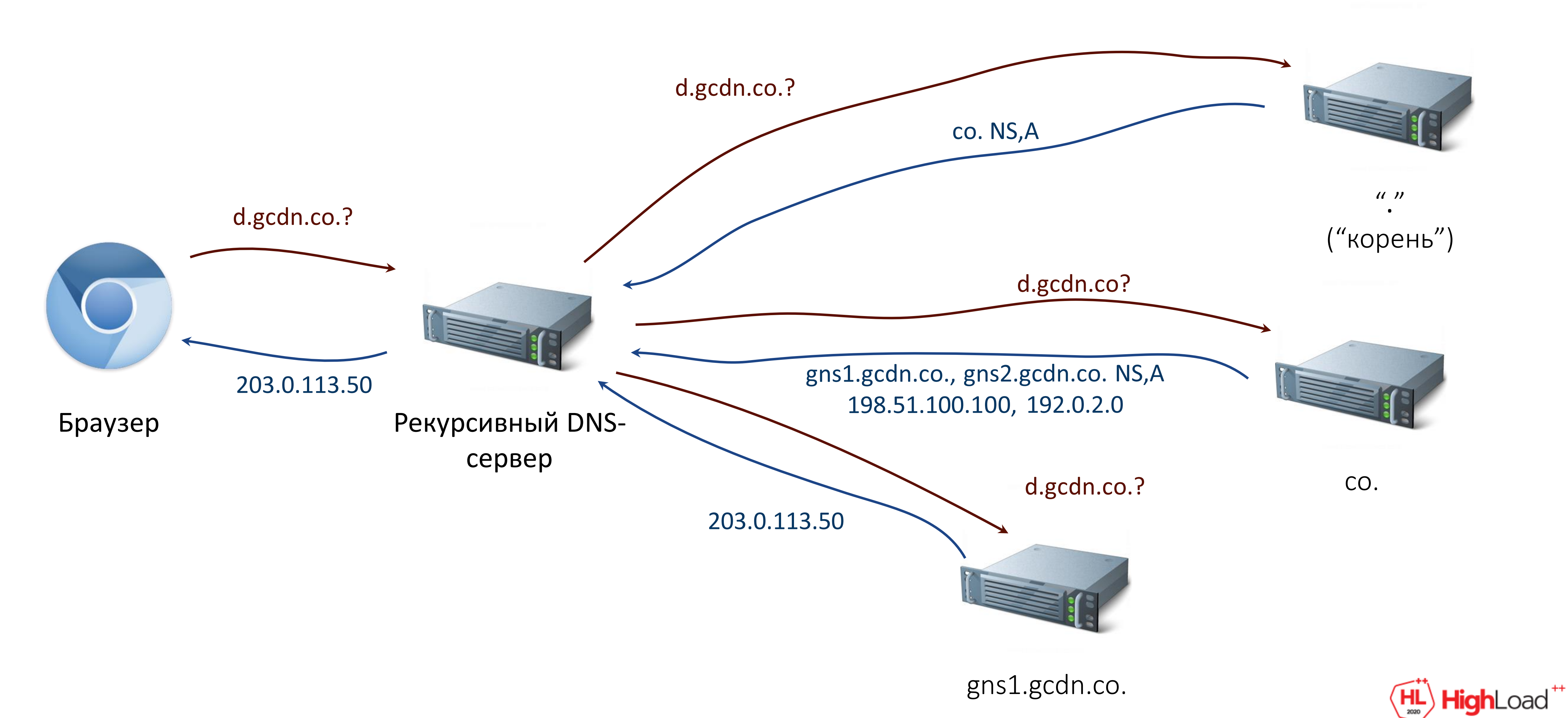
- Получает IP-адрес сервера через DNS — **latency!**
- Выполняет запрос используя IP-адрес — **latency + каналы!**



Что такое DNS или как получить IP-адрес?

- DNS:
  - Рекурсивный кеширующий сервер
  - Рекурсивные запросы




d.gcdn.co.?
co. NS,A
d.gcdn.co.?
“.”
(“корень”)
d.gcdn.co?
gns1.gcdn.co., gns2.gcdn.co. NS,A
198.51.100.100, 192.0.2.0
203.0.113.50
Браузер
Рекурсивный DNS-
сервер
d.gcdn.co.?
co.
203.0.113.50
gns1.gcdn.co.
HL 2020
HighLoad++

Что такое DNS или как получить IP-адрес?

- DNS:
  - Рекурсивный кеширующий сервер
  - Рекурсивные запросы
  - кэш и TTL
  - **Авторитетные DNS-серверы** <- будем говорить об этом



Что такое DNS или как получить IP-адрес?

- DNS:
  - Рекурсивный кеширующий сервер
  - Рекурсивные запросы
  - кэш и TTL
  - **Авторитетные DNS-серверы** <- будем говорить об этом
    - IP networks
      - BGP routing
        - Autonomous systems



Как работает CDN. Кратко

- Есть web-сервер с источником данных
- Клиент прописывает CNAME-запись с определённым значением



CDN-ресурс

cdn.example.com IN CNAME d.gcdn.co



Как работает CDN. Кратко

- Есть web-сервер с источником данных (origin)
- Клиент прописывает CNAME-запись с определённым значением
- В url указывает CDN-домен
- Мы проксируем трафик к origin'у и кешируем ответы
- Profit



Как работает CDN. Кратко

- Есть web-сервер с источником данных (origin)
- Клиент прописывает CNAME-запись с определённым значением
- В url указывает CDN-домен
- Мы проксируем трафик к origin'y и кешируем ответы
- Profit
- Предоставляли красивый поддомен в своём домене



браузер

CDN Edge

Origin



BIND и GeoIP

- Эталонный DNS-сервер
- Работает
- Есть механизм view (split horizon) по geoip-признаку
- Доставляем конфиг puppet’ом
- Неудобный конфиг



"Geodns" dns server in GO

- "This is the DNS server powering the NTP Pool system and other similar services"
- Написан на Go, простой конфиг зоны в json
- Делает то что нужно: ответ на основе базы geoip
- Есть тип записи Alias
- Конфиг получаем по cron



```
% cat d.gcdn.co.json
{
  "zonename": "d.gcdn.co",
  ...
  "ttl": 300,
  "data": {
    "": {"a": [["192.168.1.3",70], ["172.16.10.3",30]]},
    "ru": {"a": [["192.168.20.3", 20], ["172.16.40.5", 20]]},
    "en": {"a": [[" 192.0.2.2", 20], ["172.16.40.5", 80]]},
    "asia": {"a": [["203.0.113.1",20], ["172.16.40.5",20]]},
    "fi": {"alias": ["ru"]},
    ...
```



Кастомизация cdn-ресурсов клиентов

- ограничения на терминацию трафика:
  - регуляторы (например, трансляция матчей)
  - снижение стоимости услуги
- Отдельные зоны клиента приходится кастомизировать либо по запросу клиента, либо оперативно чтобы "не перелить"
- За 2я DNS-запросами может быть 2 миллиона http-запросов



Кастомизация cdn-ресурсов клиентов

- ограничения на терминацию трафика:
  - регуляторы (например, трансляция матчей)
  - снижение стоимости услуги
- Отдельные зоны клиента приходится кастомизировать либо по запросу клиента, либо оперативно чтобы "не перелить"
- За 2я DNS-запросами может быть 2 миллиона http-запросов
- Ввели компонент балансировщика для формирования зон



Anycast

1.2.3.4

1.2.3.4

1.2.3.4

1.2.3.4

1.2.3.4

1.2.3.4



Anycast

- В большом интернете — минимум /24 подсеть
- BGP (Border Gateway Protocol)
- Своя автономная система
- BGP ничего не знает про latency
- Базово используется as path
- Количество пиров
- Про уровни провайдеров (tier).



IPv4 Adjacencies		
ASN	**Name**	**Count**
AS6939	Hurricane Electric LLC	9,070
AS24482	SG.GS	6,623
AS174	Cogent Communications	6,403
AS3356	Level 3 Parent, LLC	6,062
AS36236	NetActuate, Inc	5,733
AS3549	Level 3 Communications, Inc. (GBLX)	5,667
AS199524	G-Core Labs S.A.	5,524
AS14840	BR.Digital Provider	4,610
AS51185	Onecom Global Communications LTD	4,226
AS57463	NetIX Communications Ltd.	4,036

https://bgp.he.net/report/peers




HL
2020
HighLoad++

Domain not found. Регистратор

- Регистратор разделегировал наш домен по жалобе
- Вводим категории клиентов
- Переходим к регистратору с расширенной юридической поддержкой



Domain not found. DNS-сервер

- Иногда долго отвечает или дропает пакеты
- Мало внутренних метрик, нет статистики запросов
- Слабая поддержка со стороны разработчика, свои патчи
- DNS flag day



GeoDNS: local answer time(sum of 5 local dns responses)

2.5 s
2.0 s
1.5 s
1.0 s
500 ms
0 ms

08:20 08:30 08:40 08:50 09:00 09:10 09:20 09:30 09:40 09:50 10:00 10:10 10:20 10:30 10:40 10:50 11:00 11:10

GeoDNS: udp port 53 socket receive queue size

3.0 Mil
2.5 Mil
2.0 Mil
1.5 Mil
1.0 Mil
500 K
0

08:20 08:30 08:40 08:50 09:00 09:10 09:20 09:30 09:40 09:50 10:00 10:10 10:20 10:30 10:40 10:50 11:00 11:10

ix
ix
co
de
ar
fr
fr
w
ts
ti



Domain not found. Anycast per NS server

- Разные точки присутствия для анонса подсетей
- Авария в одном ДЦ не ломает разрешение имён в регионе



Хрупкость интернета

- jitter
- packet loss



Хрупкость интернета

- jitter
- packet loss
- Сделали доставку файлов зон через P2P

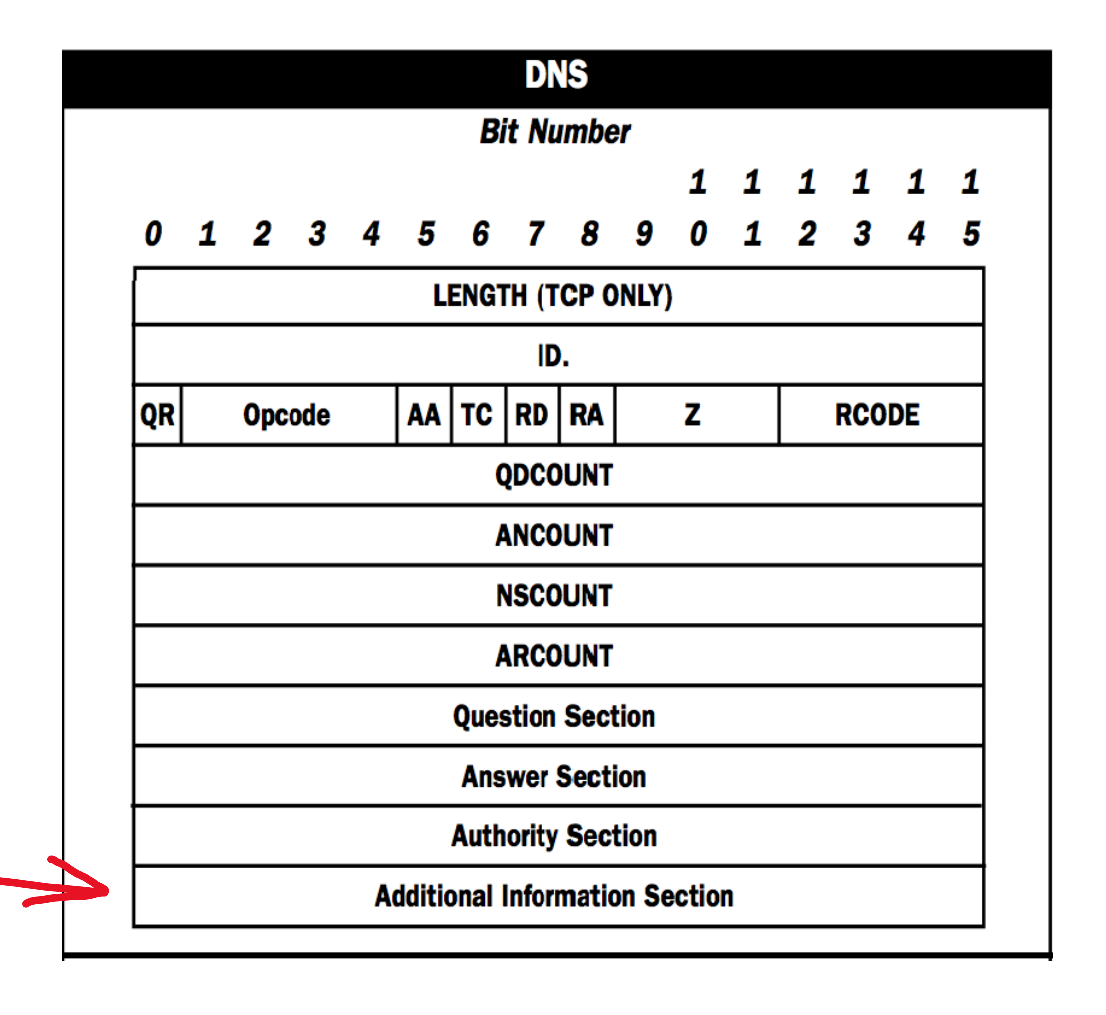


Проблемы GeoIP

- Мы видим IP-адрес рекурсора, не конечного пользователя
- Есть EDNS
- Запросов с EDNS — 30% queries (google public dns — yes, cloudflare — no, quad9 — yes)
- Не полное покрытие
- Сетевая топология не соответствует географии
- eventually inconsistent
- если адреса нет в geoip — используем гео данные DNS-сервера



EDNS Client subnet





Стратегии управления трафиком

- Отправить всё на anycast-адрес
  - Не гарантирует latency
  - Слабый контроль перегрузки со стороны северов
- Отправить на unicast-адреса
  - Может страдать latency
  - Неточность базы geoip
  - Время фэйловера в случае полного отказа ДЦ



Стратегия управления трафиком. Гибрид

- Часть на unicast-адреса
- Часть трафика на anycast-адреса:
  - Страна
  - Континент
  - Мировой anycast



DNS. Точки отказа

- Top level domain (TLD)
- NS-серверы — в случае DNS-хостинга
- Разные /24 сети в разных точках присутствия
- Автономные системы — особенный случай и требования для TLD



DNS Metrics

- prometheus наше всё, scrape, edns
- парсинг метрик geodns
- Dnstap — детальная статистика запросов, но "geodns" не может



dnstap-enabled DNS server

nameserver process

DNS worker thread

dnstap payload

circular queue

dnstap payload

dnstap I/O thread

DNS worker thread

dnstap payload

circular queue

dnstap payload

DNS queries

DNS responses

byte stream socket

dnstap capture process

dnstap receiver

dnstap payload

dnstap log file



Query stats

- пишем напрямую в clickhouse (без kafka, certbased auth)
- 20гб в сутки
- собираем через dnstap
- есть буфер
- потеряли так потеряли



Свой бекенд coredns

- сравнили с альтернативами
- понравился PowerDNS Auth (бекенды, lua), он на C++



Java

I am no longer using this memory

Memory

So I'm just going to recycle it

Memory

Recycle

C++

@System32Comics

Memory

Pick it up bitch

I ♥ Coding

Memory

@System32Comics



Свой бекенд coredns

- сравнили с альтернативами
- понравился PowerDNS Auth (бекенды, lua), он на C++
- команда go-разработчиков
- плагины, бекенды, tracing, dnstap, rfc compliance
- портирование geodns со своими патчами



Тестирование DNS-сервера

- flamethrower (нагрузочная “стрелялка”)
- respdiff (сравнение ответов разных DNS-серверов)
- dig, bash и “интеллектуальная собственность” в drone ci
- perf, flamescope при анализе проблем производительности



FLAMESCOPE Home

← Back

Rows 50 Enhanced

0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115 120 125 130 135 140 145 150 155 160 165 170 175 1



Давайте продавать наш DNS

- Между CDN balancer и dns-серверами добавили DNS API
- Поняли что у клиентов совершенно другие запросы
- Пока нет клиентов DNSaaS с 80 точками присутствия



Location: World

Type: Raw Performance | Resolver Simulation | Uptime | Quality

Period: Last 30 days

	DNS name	Query Speed
1	WordPress.com	15.38 ms
2	Cloudflare	16.54 ms
3	Gransy AnycastDNS	16.65 ms
4	dnsimple	19.22 ms
5	DigitalOcean	21.66 ms
6	G-Core	22.36 ms
7	NS1	23.4 ms
8	DNSMadeEasy	24.97 ms
9	Constellix	25 ms
10	Rage4	26.33 ms
11	UltraDNS	27.33 ms
12	Zilore	29.72 ms
13	Verizon ROUTE	30.92 ms
14	Route53	33.16 ms



Будущее

- Использование данных BGP для активных проверок
- Эффективная утилизация каналов между IX/transit через DNS
- "откуда -> куда", а "тут могут принять такой-то трафик"



Ворох проблем

- опасность каскадного отключения
- провайдер может использовать несколько апстримных dns-серверов в разных регионах
- некорректная балансировка — постоянное обновление “карты”
- DPI может проверять ip-адрес по SNI



Вопросы?

- konstantin at neuroops.link
- https://t.me/kostya_keeper
- https://twitter.com/clickfreakbit/

